Predicting asthma attacks in primary care: protocol for developing a machine learning-based prediction model

BMJ Open

Predicting asthma attacks in primary care: protocol for developing a machine learning-based prediction model

Holly Tibble,[1,2] Athanasios Tsanas,[1,2] Elsie Horne,[1,2] Robert Horne,[2,3] Mehrdad Mizani,[1,2] Colin R Simpson,[2,4] Aziz Sheikh[1,2]

To cite: Tibble H, Tsanas A, Horne E, *et al*. Predicting asthma attacks in primary care: protocol for developing a machine learning-based prediction model. *BMJ Open* 2019;**9**:e028375. doi:10.1136/bmjopen-2018-028375

► Prepublication history and additional material for this paper are available online. To view these files, please visit the journal online (http://dx.doi.org/10.1136/bmjopen-2018-028375).

Received 7 December 2018
Revised 2 April 2019
Accepted 4 June 2019

[1]Usher Institute of Population Health Sciences and Informatics, Edinburgh Medical School, College of Medicine and Veterinary Medicine, University of Edinburgh, Edinburgh, UK
[2]Asthma UK Centre for Applied Research, Edinburgh, UK
[3]University College London, London, UK
[4]School of Health, Victoria University of Wellington, Wellington, UK

Correspondence to
Holly Tibble;
holly.tibble@ed.ac.uk

ABSTRACT

Introduction Asthma is a long-term condition with rapid onset worsening of symptoms ('attacks') which can be unpredictable and may prove fatal. Models predicting asthma attacks require high sensitivity to minimise mortality risk, and high specificity to avoid unnecessary prescribing of preventative medications that carry an associated risk of adverse events. We aim to create a risk score to predict asthma attacks in primary care using a statistical learning approach trained on routinely collected electronic health record data.

Methods and analysis We will employ machine-learning classifiers (naïve Bayes, support vector machines, and random forests) to create an asthma attack risk prediction model, using the Asthma Learning Health System (ALHS) study patient registry comprising 500 000 individuals across 75 Scottish general practices, with linked longitudinal primary care prescribing records, primary care Read codes, accident and emergency records, hospital admissions and deaths. Models will be compared on a partition of the dataset reserved for validation, and the final model will be tested in both an unseen partition of the derivation dataset and an external dataset from the Seasonal Influenza Vaccination Effectiveness II (SIVE II) study.

Ethics and dissemination Permissions for the ALHS project were obtained from the South East Scotland Research Ethics Committee 02 [16/SS/0130] and the Public Benefit and Privacy Panel for Health and Social Care (1516–0489). Permissions for the SIVE II project were obtained from the Privacy Advisory Committee (National Services NHS Scotland) [68/14] and the National Research Ethics Committee West Midlands–Edgbaston [15/WM/0035]. The subsequent research paper will be submitted for publication to a peer-reviewed journal and code scripts used for all components of the data cleaning, compiling, and analysis will be made available in the open source GitHub website (https://github.com/hollytibble).

Strengths and limitations of this study

- This analysis is based on a large, representative dataset comprising over 500 000 individuals recruited from 75 general practices across Scotland.
- We will employ novel applications of established machine learning and training data enrichment methodologies.
- The prediction model we develop will be tested in unseen large external dataset, namely the SIVE II dataset.
- This derivation and validation work will be undertaken in NHS Scotland; there will therefore be a need for further validation work in other UK nations and international contexts.

INTRODUCTION

Asthma is a long-term lung disease characterised by inflammation of the airways, which may manifest as episodic wheezing, chest tightness, coughing, and shortness of breath. An asthma attack is the sudden worsening of symptoms, which may prove fatal.[1] In 2017, asthma was estimated to affect 235 million people worldwide.[2] In 2015 alone, 1434 people died from asthma attacks in the UK—a rate of 2.21 deaths per 100 000 person-years.[3] Asthma attack incidence is reported to be between 0.01 and 0.78 events per person-year, depending on the definition of attacks and the population (eg, primary care, secondary care).[4–6]

Asthma therapy typically follows a fairly linear path—beginning with a short-acting bronchodilator in the individuals without persistent asthma symptoms and adding preventative treatments and long-acting bronchodilators in the individuals with more persistent asthma symptoms.[7 8] The individuals with persistent troublesome symptoms and/or considered to be at very high risk may be prescribed biologicals and/or oral steroids.[9] Oral steroids are often considered a last resort due to their undesirable safety profile including increased risk of diabetes,[10–12] osteoporosis,[13–15] and affective and psychotic disorders.[15–18]

It follows that the determination of those at high risk for asthma attacks is crucial in order to prevent attacks and minimise the risk of unnecessary side effects. Furthermore, the 2014 National Review of Asthma

Deaths found that 45% of asthma deaths in the study year occurred without the patient requesting medical help or before help could be provided.[5] Increased awareness of the risk could prevent the patients with asthma from delay in seeking medical care and preventing fatality.

While it might seem intuitive that the patients with most severe daily symptoms exhibit greater risk of severe morbidity and mortality, research suggests that these symptoms may be a suboptimal clinical marker of asthma attack risk.[19] Indeed, some people with asthma are more prone to asthma attacks than others, with asthma attack history being the strongest risk factor for future asthma attacks.[20–23] Other commonly identified risk factors for asthma attacks include poor asthma control[24–27] (often a result of poor adherence to preventative therapy[28–31]), smoking,[24 27 32–34] history of hospital admission,[21 24] history of oral steroid use,[24] obesity,[27 34–38] access to medicines,[39 40] socioeconomic status,[41 42] and viral respiratory infections.[43–45]

Despite the identification of many risk factors, identifying high-risk individuals has proven a challenging task. Logistic regression, the most commonly used statistical method for event prediction, is known to predict outcomes poorly when there is *class imbalance* (event and no event),[46] and we expect the problem investigated in this study assessing asthma attacks will be highly imbalanced. For example, a model could predict that a very rare event would never occur, and it would be correct in the vast majority of cases. As such, most prediction models report high *specificity* (correctly predicting low attack risk to those who did not have attacks), but low *sensitivity* (correctly predicting high risk in those who did go on to have attacks),[4 24 41 47–51] which results in less reliable risk prediction for patients at high risk.

In a recent study by Finkelstein and Jeong,[52] sensitivity (and specificity) in excess of 75% was achieved for all classifiers (adaptive Bayesian network, naïve Bayes classifier, and support vector machine) predicting asthma attacks a week in advance using a sample of just over 7000 records of home tele-monitoring data. They found substantial improvements in model sensitivity using *training enrichment* methods, pre-processing the training data to improve the performance in the testing data—for example, by increasing the prevalence of the rare outcome in the training data to balance the classes.

RESEARCH AIM

We aim to create a personalised risk assessment tool to assist primary care clinicians in predicting asthma attacks over a period of 1, 4, 12, 26, and 52 weeks, employing machine-learning methodologies such as naïve Bayes classifiers, random forests, and support vector machines, as well as ensemble algorithms. The model will build on previous research[4 24 41 47–52] to improve the sensitivity of our event prediction, without unduly compromising the specificity. This is crucial in order to reduce prescribing steroid, diminish the long-term effects of high steroid use over a lifetime, which have adverse effects,[10–18] and reduce patient anxiety when risk of an asthma attack is low.

Primary care consultations provide the opportunity for patients and clinicians to assess changes to asthma attack risk, which can be used to promote patients to seek emergency care if there is a significant deterioration in their symptoms and to promote risk-reducing lifestyle choices.

METHODS

Data sources and permissions

The derivation dataset used for training, validating, and testing the model will be the Asthma Learning Healthcare System (ALHS) dataset, created in order to develop and validate a prototype learning health system for asthma patients in Scotland.[53] The ALHS study aims to increase understanding of variation in asthma outcomes and create benchmarks for clinical practice in order to reduce suboptimal care by repurposing patient data to create a continuous loop of knowledge generation, evidence-based clinical practice change, and change assessment. The study dataset contains patient demographics from the patient registry, primary care prescribing records, primary care encounters, Accident and Emergency (A&E) records, hospital inpatient admissions and deaths, linkable by an anonymised unique identifier. Datasets were extracted between November 2017 and August 2018 for the period January 2000 to December 2017, as shown

Table 1 Metadata for clinical data sources in derivation dataset (ALHS)

Data Source	Number of Records	Number of Individuals	Extraction Date	Data Specification Date Range
Primary Care Prescribing*	4 709 231	47 095	October 2018	January 2009–April 2017
Primary Care Encounters*	11 766 100	49 307	March 2018	January 2000–November 2017
Accident & Emergency	1 831 789	500 321	November 2017	June 2007–September 2017
Hospital Inpatient Admissions	1 668 957	342 838	August 2018	January 2000–March 2017
Mortality	*NA*	91 758	May 2018	January 2000–March 2017

*Records available for subset of study population with asthma diagnosis only.

Table 2 Metadata for clinical data sources in external dataset (SIVE II)

Data Source	Number of Records	Number of Individuals	Extraction Date	Data Specification Date Range
Primary Care Prescribing	29 360 448	1 073 377	May 2017	January 2003–March 2017
Primary Care Encounters	31 878 423	1 887 957	May 2017	January 2000*–March 2017
Accident & Emergency	4 116 561	1 247 314	April 2017	June 2007– August 2016
Hospital Inpatient Admissions	3 549 174	794 937	April 2017	January 2000–March 2017
Mortality	*NA*	215 466	April 2017	January 2000– March 2017

*Diagnosis codes entered in this period, but post-dated from 1940 onwards retained.

in table 1, along with the number of records and unique individuals before data cleaning.

In order to verify that the prediction model performance is not limited to the development dataset and that it generalises well in new, unseen data presented to the classifier in the training process, we will evaluate its performance using an external cohort study dataset, the second Seasonal Influenza Vaccination Effectiveness (SIVE II) cohort study,[54 55] which used a large national primary care (1.25 million individuals from 230 Scottish general practices) and laboratory-linked dataset to evaluate live attenuated and trivalent inactivated influenza vaccination effectiveness. The SIVE II dataset contains records from the same sources (primary and secondary care) and modalities (diagnosis and date) as the ALHS dataset (extraction and specification dates are shown in table 2), and can be harmonised such that variables and value sets are aligned. In Appendix A, we detail the data harmonisation plan, that is, we list the key variables to be used in the following analyses, their format in each dataset (for example, whether age is pre-coded into 5-year bands), and the common denominator format that will be used in the analyses to ensure the highest degree of concordance during the validation stage.

Patient and public involvement

This analysis plan was constructed with the assistance of the Asthma UK Centre for Applied Research (AUKCAR) Patient and Public Involvement (PPI) group. The particular importance of avoiding a substantial decrease in specificity in order to gain higher sensitivity was a result of discussions within this group about the burden of side effects from preventative treatment.

Inclusion criteria

We will identify our study population as all adults (aged 18 and over) with asthma being identified by clinical diagnoses (Read codes), without a chronic obstructive pulmonary disease (COPD) diagnosis, and with relevant prescribing records in primary care. Patients with missing sex or age information will be removed; this and any other patient exclusions from further analysis will be explicitly detailed.

All records from the derivation dataset (ALHS) will be left-censored on January 2009 in order to align with the primary care prescribing data and right-censored on March 2017 in order to align with the mortality, primary care, and inpatient hospital admission records, as presented in table 1. Similarly, records from the external dataset (SIVE II) will be left-censored on January 2003 in order to align with the primary care prescribing data and right-censored on August 2016 to align with the A&E records, as shown in table 2. There is a high probability that some individuals will have been recruited into both studies, and therefore those individuals will be flagged in the external testing dataset and removed from the study pool.

Outcome ascertainment

We will identify asthma attacks, defined by the American Thoracic Society/European Respiratory Society,[56] as a prescription of oral corticosteroids, an asthma-related A&E visit, or an asthma-related hospital admission. Additionally, deaths occurring with asthma as the primary cause will be labelled as asthma attacks. Instances of multiple attack indicators occurring within a 14-day period were coded as a single attack.

Patient characteristics, confounders, and missing data

Patient characteristics at baseline will be reported and included as time-varying confounders in analyses. For all characteristics derived from Read codes, full code lists will be provided as online supplementary materials.

Demographics: Age, sex, rurality, and social deprivation will be extracted from the primary care registry. Social deprivation is measured using quintiles of the Scottish Index of Multiple Deprivation (SIMD), a geographic measure derived using data on income, employment, education, health, access to services, crime, and housing.[57] Rurality is defined using the Scottish Government Urban Rural Classification Scale (6-fold scale).[58] While missing age and/or sex are exclusion criteria for the study sample, absence for rurality and social deprivation will be coded as 'missing.'

Practice Location: Practice location will be included in order to account for clustering of patients by region. Location will be coded using the Nomenclature of Territorial Units for Statistics[59] (NUTS 3) codes, linked from the registered practice data zone (2001).

Asthma Severity: Asthma severity will be categorised using the British Thoracic Society's 2016 5-step treatment classification.[60] Severity will be considered time-dependent and will be determined using prescribing records at any change in regimen.

Smoking Status: Smoking status will be derived from primary care data and presented as a 3-level variable, namely current, former, and non-smoker, using the most recent smoking Read code at any day. Smoking status will be considered time-dependent and determined using the most recent Read code records, and the individuals with unknown smoking status will be coded as non-smokers.[61 62]

Blood Eosinophil Count: Blood eosinophil count will be derived from primary care Read codes and will be dichotomised at ≥400 cells/μL. The individuals with non-recorded blood eosinophil count will be coded as missing. Blood eosinophil count will be considered time-dependent and determined using the most recent Read code record.

Obesity: Obesity will be derived from body mass index (BMI) or height and weight records in primary care data and will be presented as a binary variable (BMI≥30). The individuals with unknown BMI will be coded as non-obese. Obesity will be considered time-dependent and determined using the most recent Read code record.

Comorbidity: Comorbidity will be defined by 17 dichotomous (unweighted) variables representing the diagnostic categories of the adapted Charlson Comorbidity Index.[63 64] Additionally, active diagnoses of rhinitis, eczema, gastro-oesophageal reflux disease, nasal polyps, and anaphylaxis will be recorded; all identified by Blakey *et al* as contributing characteristics to increased asthma attack risk.[65] Comorbidities will be considered time-dependent and determined using all prior Read code records.

Previous Healthcare Usage: The number of repeat prescriptions of preventer medication and the number of primary care asthma encounters (days on which at least one asthma related code was recorded) in the previous year will be derived from primary care prescribing and Read code records, respectively. Both will be considered time-dependent and determined using records from the previous calendar year.

Asthma Control: The mean short-acting beta-2 agonist dose per day will be estimated retroactively by examining the dates between prescriptions. The most recent peak expiratory flow measurement at any time will be recorded (categorical, based on percentage of previous maximum) or coded as missing if that measurement was more than 7 days ago. Adherence to preventer therapy will be approximated using the medication possession ratio,[66] calculated from primary care prescribing records.

History of Asthma Attacks: Prior asthma attacks will be identified solely using primary care prescribing records and Read codes. This is because primary care practitioners will not be able to make use of secondary care records when utilising this risk score with patients. Both the prior number of attacks and the time since the last attack will be included as predictors and will be considered time-dependent and accurate at the weekly level.

Analysis plan

The derivation dataset (ALHS) will be divided into three partitions: 60% for training, 20% for model comparison (validation), and 20% to assess performance (testing). In our training subset, the first partition, we will train machine learning models (classifiers) with varying hyperparameters, predicting asthma attack occurrence in the following 1, 4, 26, and 52 weeks. We will run 100 iterations for statistical confidence, each time randomly permuting samples prior to determining the three subsets. The *no free lunch theorem* in machine learning suggests that there is no classifier (or more generically a machine learning tool) which will consistently outperform competing approaches across all settings.[67] Therefore, given that we do not know a priori which classifier will work best in this application, we will apply naïve Bayes classifiers for benchmarking and then employ more advanced state-of-the-art principled supervised learning algorithmic tools such as support vector machines, random forests, and ensembles (classifier combinations) to investigate which algorithm leads to more accurate results.

A selection of *training enrichment* methods will be trialled in order to assess how to best overcome poor performance as a result of low outcome prevalence. Typically, modelling rare events results in reduced sensitivity (the proportion of the individuals who had attacks that were detected), so the individuals predicted to be at low risk will have a high rate of asthma attacks. As such, the start of this process (the first 20 iterations of training each model) will be repeated five times using:

1. the original analysis dataset,
2. original data with additional duplicates of the positive outcome records (a method known as *over-sampling*) ,[68]
3. original data, with a selection of the negative outcome records removed (*under-sampling*),[68]
4. original data with additional slightly modified duplicates of the positive outcome records, with a selection of the negative outcome records removed (*synthetic minority over-sampling; SMOTE*),[68 69]
5. original data, using the outcome classification threshold to maximise the primary metric—the Matthew's correlation coefficient (MCC)[70]—identified using golden-section search optimisation.[71]

By assessing the average performance by classification method class, in each set of iterations, we will determine which enrichment method is the most appropriate overall for the data and to be continued accordingly.

In the validation partition, with all 100 iterations for the selected enrichment methods, we will identify the highest performing model as that with the highest mean MCC across iterations; in the event of a tie, the model with the highest iteration-minimum MCC will be selected.

Model testing will be conducted on the selected model (figure 1) in the derivation testing partitions. Model calibration will be assessed by comparing observed rate of

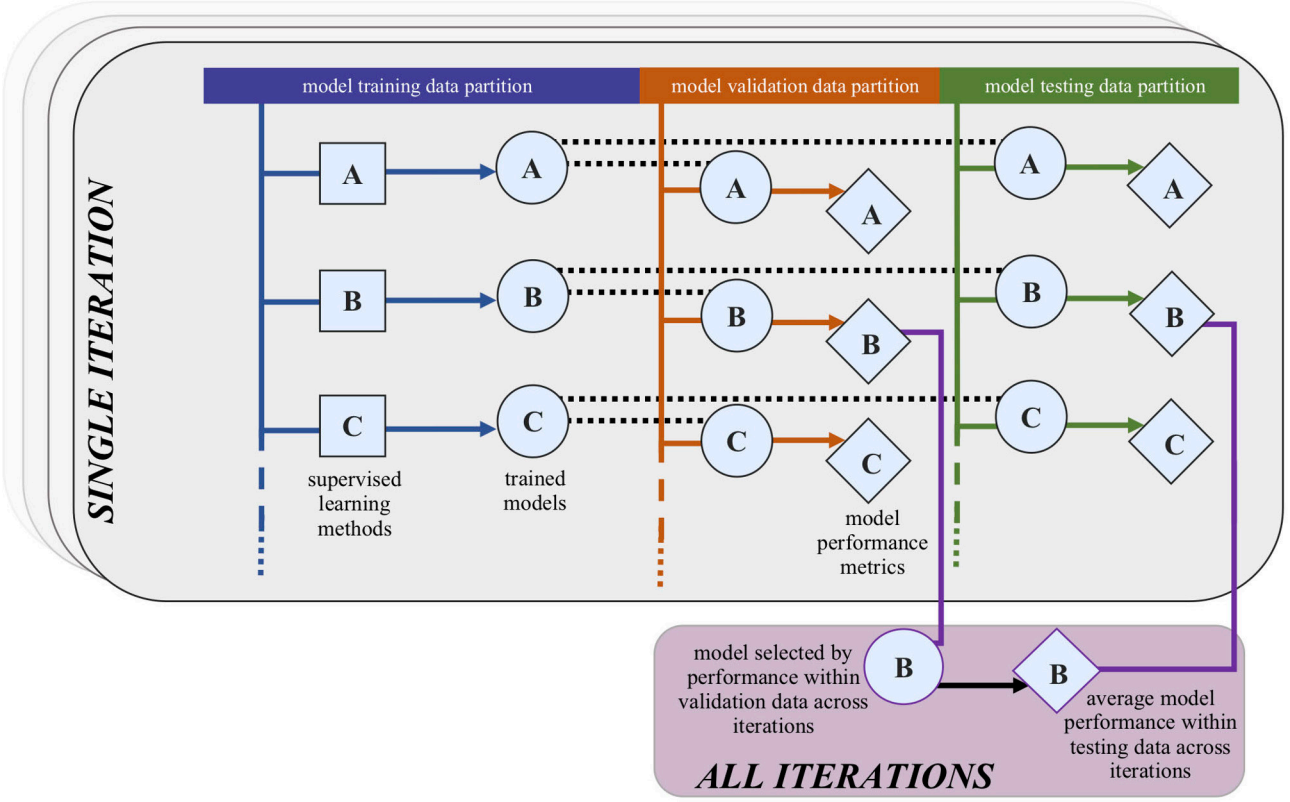

Figure 1 Process of selecting the highest performing model from the validation data and the average performance of this model across iterations in the testing dataset. In the foreground, we have the first iteration. We will use 100 iterations for statistical confidence, randomly permuting the data into training, validation, and testing subsets in each iteration.

incidence by predicted risk for the full population and by exhaustive population subgroups, including asthma severity, prior number of asthma attacks, age, and smoking status (particularly useful to assess possible contamination by asthma-COPD overlap syndrome (ACOS)). We will also check the calibration between the predicted risk and the attack incidence, stratified by the source of the asthma attack record (in primary care, A&E presentation, or inpatient admission). Performance in the testing datasets will be assessed using the MCC, and the additional metrics of sensitivity, specificity, positive and negative predictive values, and the F_1 measure,[72] along with information criteria such as the Bayesian Information Criterion are calculated to obtain a trade-off between model complexity and accuracy. Confusion matrices (also known as contingency tables) will be made available as online supplementary materials.

The derivation dataset will be re-used in its entirety to retrain the model based on the final classifier and hyperparameter selection. Model testing will then be conducted in the external dataset, which consists of data unseen in the model derivation, using this trained model. Distributions of predictors between the derivation and external datasets will be assessed (indirectly) to contextualise the generalisability findings. The aforementioned metrics will be reported.

Finally, we will re-train the model using the hyperparameter specifications from the best performing model, with a modified version of the derivation dataset which incorporates data extracted from secondary care records (such as A&E presentations for asthma attack not captured in primary care records) in the determination of the risk factors. This allows us to evaluate the added value of secondary care data linkage in the prediction of impending asthma attacks and will be determined by the same metrics used for the primary model evaluation.

All analyses will be conducted in R (though the RStudio interface), and details on the functions, the hyperparameter within each classifier, and the ranges assessed herein are provided in Appendix B.

Ethics and dissemination

All authors with data access have completed the Safe Users of Research data Environment training, provided by the Administrative Data Research Network. All analysis will be conducted in concordance with the National Services Scotland Electronic Data Research and Innovation Service (eDRIS) user agreement. This study protocol will be registered with the European Union electronic Register of Post-Authorisation Studies (EU PAS Register) as a non-interventional post-authorisation study (PAS) before any data analysis is initiated.

The subsequent research paper will be submitted for publication in a peer-reviewed journal and will be written in accordance with TRIPOD: *transparent reporting of a multivariable prediction model for individual prognosis*

or diagnosis[73] and RECORD: *reporting of studies conducted using observational routinely-collected health data*[74] guidelines. Code scripts used for all components of the data cleaning, compiling, and analysis will be made available in the open source GitHub website at https://github.com/hollytibble.

A lay summary of this protocol paper, and the subsequent research results paper, will be made available online (via an open source platform) in order to heighten the impact and accessibility of this work. A lay summary on asthma will be provided as online supplementary materials.

CONCLUSIONS

This project will further advance asthma attack risk prediction modelling and will inform on the future direction of routine data linkage in Scotland, which is likely to have additional benefits for other health systems in the UK and internationally.

Acknowledgements The authors would like to thank Eleftheria Vasileiou, Amy Tilbrook, Mome Mukherjee, and Jill Tibble for their contributions to the analysis plan, data management, and proof-reading of this manuscript, and the Asthma UK Centre for Applied Research (AUKCAR) Patient and Public Involvement group for their contribution to the analysis plan.

Contributors HT and AT conceived and planned the analysis. HT and RH specified the medication adherence measures. HT, EH, CS, MM, and AS constructed the covariate (and associated Read Coding) lists for the model. HT wrote the first draft, with contributions from all authors. All authors (HT, AT, EH, RH, MM, CRS, and AS) approved the final version and jointly take responsibility for the decision to submit this manuscript to be considered for publication.

Funding HT is supported by College of Medicine and Veterinary Medicine PhD (eHERC/Farr Institute) Studentships from The University of Edinburgh. EH is supported by a Medical Research Council PhD Studentship (eHERC/Farr). MAM's Newton International Fellowship is awarded by the Academy of Medical Sciences and Newton Fund. This work is carried out with the support of the Asthma UK Centre for Applied Research [AUK-AC-2012-01] and Health Data Research UK, an initiative funded by UK Research and Innovation Councils, National Institute for Health Research (England) and the UK devolved administrations, and leading medical research charities. The ALHS dataset was created with funding from the National Environment Research Council [NE/P011012/1]. The SIVE II dataset was created with funding from the National Institute for Health Research (NIHR) Health Technology Assessment programme [13/34/14]—the views and opinions expressed therein are those of the authors and do not necessarily reflect those of the Health Technology Assessment programme, NIHR, NHS, or the Department of Health.

Competing interests None declared.

Patient consent for publication Not required.

Ethics approval Permissions for the ALHS project were obtained from the South East Scotland Research Ethics Committee 02 [16 /SS/0130] and the Public Benefit and Privacy Panel for Health and Social Care (1516 – 0489) . Permissions for the SIVE II project were obtained from the Privacy Advisory Committee (National Services NHS Scotland) [68/14] and the National Research Ethics Committee West Midlands - Edgbaston [15/WM/0035].

Provenance and peer review Not commissioned; externally peer reviewed.

REFERENCES

1. Bush A, Griffiths C. Improving treatment of asthma attacks in children. *BMJ* 2017;359:j5763.
2. World Health Organisation. *Asthma Fact Sheet (2017). World Health Organisation Fact Sheets*: World Health Organization, 2017.
3. Asthma UK. *UK asthma death rates among worst in Europe*, 2017.
4. Loymans RJB, Debray TPA, Honkoop PJ, *et al*. Exacerbations in adults with asthma: a systematic review and external validation of prediction models. *J Allergy Clin Immunol Pract* 2018;6.
5. *Royal College of Physcians*. Why asthma still kills: The National Review of Asthma Deaths (NRAD), 2014.
6. Mukherjee M, Nwaru BI, Soyiri I, *et al*. High health gain patients with asthma: a cross-sectional study analysing national Scottish data sets. *NPJ Prim Care Respir Med* 2018;28:27.
7. British Thoracic Society & Scottish Intercollegiate Guidelines Network. British guideline on the management of asthma. *SIGN Guidel* 2014.
8. Currie GP, Douglas JG, Heaney LG. Difficult to treat asthma in adults. *BMJ* 2009;338:b494.
9. British Thoracic Society, Research Unit of the Royal College of Physicians of London, King's Fund Centre & National Asthma Campaign. Guidelines For Management Of Asthma In Adults: I: Chronic Persistent Asthma. *Br Med J* 1990;301:651–3.
10. Kim SY, Yoo CG, Lee CT, *et al*. Incidence and risk factors of steroid-induced diabetes in patients with respiratory disease. *J Korean Med Sci* 2011;26:264–7.
11. Suissa S, Kezouh A, Ernst P. Inhaled corticosteroids and the risks of diabetes onset and progression. *Am J Med* 2010;123:1001–6.
12. Blackburn D, Hux J, Mamdani M. Quantification of the Risk of Corticosteroid-induced Diabetes Mellitus Among the Elderly. *J Gen Intern Med* 2002;17:717–20.
13. Adinoff AD, Hollister JR. Steroid-induced fractures and bone loss in patients with asthma. *N Engl J Med* 1983;309:265–8.
14. Van Staa TP, Leufkens HG, Abenhaim L, *et al*. Use of oral corticosteroids and risk of fractures. *J Bone Min. Res* 2000.
15. Bloechliger M, Reinau D, Spoendlin J, *et al*. Adverse events profile of oral corticosteroids among asthma patients in the UK: cohort study with a nested case-control analysis. *Respir Res* 2018;19:75.
16. Dawson KL, Carter ER. A steroid-induced acute psychosis in a child with asthma. *Pediatr Pulmonol* 1998;26:362–4.
17. Kayani S, Shannon DC. Adverse behavioral effects of treatment for acute exacerbation of asthma in children: a comparison of two doses of oral steroids. *Chest* 2002;122:624–8.
18. Brown ES, Khan DA, Nejtek VA. The psychiatric side effects of corticosteroids. *Annals of Allergy, Asthma & Immunology* 1999;83:495–504.
19. Green RH, Brightling CE, McKenna S, *et al*. Asthma exacerbations and sputum eosinophil counts: a randomised controlled trial. *Lancet* 2002;360:1715–21.
20. Buelo A, McLean S, Julious S, *et al*. At-risk children with asthma (ARC): a systematic review. *Thorax* 2018;73:1–12.
21. Turner MO, Noertjojo K, Vedal S, *et al*. Risk factors for near-fatal asthma. A case-control study in hospitalized patients with asthma. *Am J Respir Crit Care Med* 1998;157:1804–9.
22. ten Brinke A, Sterk PJ, Masclee AA, *et al*. Risk factors of frequent exacerbations in difficult-to-treat asthma. *Eur Respir J* 2005;26:812-8.
23. Turner SW, Murray C, Thomas M, *et al*. Applying UK real-world primary care data to predict asthma attacks in 3776 well-characterised children: a retrospective cohort study. *npj Prim. Care Respir. Med* 2018:28, 28.
24. Loymans RJ, Honkoop PJ, Termeer EH, *et al*. Identifying patients at risk for severe exacerbations of asthma: development and external validation of a multivariable prediction model. *Thorax* 2016;71:838–46.
25. Robroeks CM, van Vliet D, Jöbsis Q, *et al*. Prediction of asthma exacerbations in children: results of a one-year prospective study. *Clin Exp Allergy* 2012;42:792–8.
26. Haselkorn T, Zeiger RS, Chipps BE, *et al*. Recent asthma exacerbations predict future exacerbations in children with severe or difficult-to-treat asthma. *J Allergy Clin Immunol* 2009;124:921–7.
27. Bateman ED, Buhl R, O'Byrne PM, *et al*. Development and validation of a novel risk score for asthma exacerbations: The risk score for exacerbations. *J Allergy Clin Immunol* 2015;135(6):e4.
28. Bårnes CB, Ulrik CS. Asthma and adherence to inhaled corticosteroids: current status and future perspectives. *Respir Care* 2015;60:455–68.
29. Fernandes AG, Souza-Machado C, Coelho RC, *et al*. Risk factors for death in patients with severe asthma. *J Bras Pneumol* 2014;40:364–72.

30. Engelkes M, Janssens HM, de Jongste JC, *et al*. Medication adherence and the risk of severe asthma exacerbations: a systematic review. *Eur Respir J* 2015;45:396–407.
31. Papi A, Ryan D, Soriano JB, *et al*. Relationship of inhaled corticosteroid adherence to asthma exacerbations in patients with moderate-to-severe asthma. *J Allergy Clin Immunol Pract* 2018;6.
32. McCarville M, Sohn MW, Oh E, *et al*. Environmental tobacco smoke and asthma exacerbations and severity: the difference between measured and reported exposure. *Arch Dis Child* 2013;98:510–4.
33. Marquette CH, *et al*. Long-term prognosis of near-fatal asthma. *Am. Rev. Respir. Dis* 1992;146:76–81.
34. Price D, Wilson AM, Chisholm A, *et al*. Predicting frequent asthma exacerbations using blood eosinophil count and other patient data routinely available in clinical practice. *J Asthma Allergy* 2016;9.
35. Black MH, Zhou H, Takayanagi M, *et al*. Increased asthma risk and asthma-related health care complications associated with childhood obesity. *Am J Epidemiol* 2013;178:1120–8.
36. Schatz M, Zeiger RS, Zhang F, *et al*. Overweight/obesity and risk of seasonal asthma exacerbations. *J Allergy Clin Immunol Pract* 2013;1:618–22.
37. Quinto KB, Zuraw BL, Poon KY, *et al*. The association of obesity and asthma severity and control in children. *J Allergy Clin Immunol* 2011;128:964–9.
38. Stingone JA, Ramirez OF, Svensson K, *et al*. Prevalence, demographics, and health outcomes of comorbid asthma and overweight in urban children. *J Asthma* 2011;48:876–85.
39. Sarpong SB, Karrison T. Sensitization to indoor allergens and the risk for asthma hospitalization in children. *Ann Allergy Asthma Immunol* 1997;79:455–9.
40. Stingone JA, Claudio L. Disparities in the use of urgent health care services among asthmatic children. *Ann Allergy Asthma Immunol* 2006;97:244–50.
41. Schatz M, Cook EF, Joshua A, *et al*. Risk factors for asthma hospitalizations in a managed care organization: development of a clinical prediction rule. *Am J Manag Care* 2003;9:538–47.
42. Rosas-Salazar C, Ramratnam SK, Brehm JM, *et al*. Parental numeracy and asthma exacerbations in Puerto Rican children. *Chest* 2013;144:92–8.
43. Bossios A, Papadopoulos N. Viruses and asthma exacerbations. *Breathe* 2006;3:51–8.
44. Jackson DJ, Johnston SL. The role of viruses in acute exacerbations of asthma. *J Allergy Clin Immunol* 2010;125:1178–87.
45. Busse WW, Lemanske RF, Gern JE. Role of viral respiratory infections in asthma and asthma exacerbations. *Lancet* 2010;376:826–34.
46. King G, Zeng L, King G. Logistic regression in rare events data. *Polit. Anal* 2001;9:137–63.
47. Lieu TA, Quesenberry CP, Sorel ME, *et al*. Computer-based models to identify high-risk children with asthma. *Am J Respir Crit Care Med* 1998;157:1173-80.
48. Smith JR, Noble MJ, Musgrave S, *et al*. The at-risk registers in severe asthma (ARRISA) study: a cluster-randomised controlled trial examining effectiveness and costs in primary care. *Thorax* 2012;67:1052–60.
49. van Vliet D, Alonso A, Rijkers G, *et al*. Prediction of asthma exacerbations in children by innovative exhaled inflammatory markers: results of a longitudinal study. *PLoS One* 2015;10:1–15.
50. Hallit S, Raherison C, Malaeb D, *et al*. Development of an asthma risk factors scale (ARFS) for risk assessment asthma screening in children. *Pediatr Neonatol* 2019;60.
51. Forno E, Fuhlbrigge A, Soto-Quirós ME, *et al*. Risk factors and predictive clinical scores for asthma exacerbations in childhood. *Chest* 2010;138:1156–65.
52. Finkelstein J, Jeong IC. Machine learning approaches to personalize early prediction of asthma exacerbations. *Ann N Y Acad Sci* 2017;1387:153–65.
53. Soyiri IN, Sheikh A, Reis S, *et al*. Improving predictive asthma algorithms with modelled environment data for Scotland: an observational cohort study protocol. *BMJ Open* 2018;8:e23289.
54. Simpson CR, Lone NI, Kavanagh K, *et al*. Evaluating the effectiveness, impact and safety of live attenuated and seasonal inactivated influenza vaccination: protocol for the Seasonal Influenza Vaccination Effectiveness II (SIVE II) study. *BMJ Open* 2017;7:e014200.
55. Simpson CR, *et al*. Seasonal Influenza Vaccination Effectiveness II (SIVE II): an observational study to evaluate live attenuated and trivalent inactivated influenza vaccination effectiveness, public health impact and safety—2010/11 to 2015/16 seasons. *Heal Technol Assess*. In Press.
56. Reddel HK, Taylor DR, Bateman ED, *et al*. An official American Thoracic Society/European Respiratory Society statement: asthma control and exacerbations: standardizing endpoints for clinical asthma trials and clinical practice. *Am J Respir Crit Care Med* 2009;180:59–99.
57. Scottish Government National Statistics Publications. *Introducing The Scottish Index of Multiple Deprivation 2016*, 2016.
58. Scottish Government. *Scottish Government Urban Rural Classification*, 2016.
59. Scottish Government. *Review of Nomenclature of Units for Territorial Statistics (NUTS) Boundaries*, 2016.
60. Society BT. British Guideline on the Management of Asthma: Quick Reference Guide. *Scottish Intercollegiate Guidelines Network* 2016.
61. Lewis JD, Brensinger C. Agreement between GPRD smoking data: a survey of general practitioners and a population-based survey. *Pharmacoepidemiol Drug Saf* 2004;13:437–41.
62. Marston L, Carpenter JR, Walters KR, *et al*. Issues in multiple imputation of missing data for large general practice clinical databases. *Pharmacoepidemiol Drug Saf* 2010;19:618–26.
63. Deyo RA, Cherkin DC, Ciol MA. Adapting a clinical comorbidity index for use with ICD-9-CM administrative databases. *J Clin Epidemiol* 1992;45:613–9.
64. Charlson ME, Pompei P, Ales KL, *et al*. A new method of classifying prognostic comorbidity in longitudinal studies: development and validation. *J Chronic Dis* 1987;40:373–83.
65. Blakey JD, Price DB, Pizzichini E, *et al*. Identifying risk of future asthma attacks using UK Medical Record Data: A Respiratory Effectiveness Group Initiative. *J Allergy Clin Immunol Pract* 2017;5:1015–24.
66. Hess LM, Raebel MA, Conner DA, *et al*. Measurement of adherence in pharmacy administrative databases: a proposal for standard definitions and preferred measures. *Ann Pharmacother* 2006;40:1280–8.
67. Wolpert DH, Macready WG. No free lunch theorems for optimization. *IEEE Trans. Evol. Comput* 1997;1:67–82.
68. He H, Garcia EA. Learning from imbalanced data. *IEEE Trans. Knowl. DATA Eng* 2009;21:1263–84.
69. Chawla N. V, Bowyer KW, Hall LO, *et al*. SMOTE: Synthetic minority over-sampling technique. *J. Artif. Intell. Res* 2002;16:321–57.
70. Chicco D. Ten quick tips for machine learning in computational biology. *BioData Min* 2017;10:1–17.
71. Kiefer J. Sequential minimax search for a maximum. *Proc. Am. Math. Soc* 1953.
72. Hripcsak G, Rothschild AS. Agreement, the F-measure, and reliability in information retrieval. *J Am Med Inform Assoc* 2005;12:296–8.
73. Collins GS, Reitsma JB, Altman DG, *et al*. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD Statement. *BMC Med* 2015;13:1.
74. Nicholls SG, Quach P, von Elm E, *et al*. The Reporting of studies conducted using observational routinely-collected health data (record) statement: Methods for arriving at consensus and developing reporting guidelines. *PLoS One* 2015;10:e0125620–3.